Paper 058-2013

The Forest and the Trees:
See it All with SAS® Visual Analytics

Nascif Abousalh-Neto, SAS Institute Inc., Cary, NC

ABSTRACT

It's a jungle out there! A data jungle, that is. With so much data to process, it's too easy to become lost. What's a data explorer to do? This paper explains how data exploration journeys usually follow a generic workflow composed of nine well-defined tasks that are easy to perform using SAS® Visual Analytics. Armed with this knowledge, you will be able to see both the forest and the trees, and never worry again about losing your way.

INTRODUCTION

See the forest for the trees (idiomatic) *To discern an overall pattern from a mass of detail; to see the big picture, or the broader, more general situation.*

The idiom "see the forest for the trees" captures at once the prize and the challenge of Business Intelligence.The fact that it is often used in the negative to indicate that someone is missing the big picture due to an abundance of details makes it an even better analogy. The need to find the "forest" while not losing sight of significant "trees" in an environment with ever increasing volume, velocity, and variety of data [Dumbill2012] has stretched the limits of data analysis. The field has been evolving over the years to keep up with these demands.

THE EVOLVING FIELD OF DATA ANALYSIS

A variety of disciplines have been developed to provide answers to the questions that data analysis tries to solve. Although these disciplines differ in their approaches and workflows, there has been some convergence as well. Figure 1 summarizes the cornerstones of the three disciplines that this paper examines: data mining, exploratory data analysis, and visual analytics.

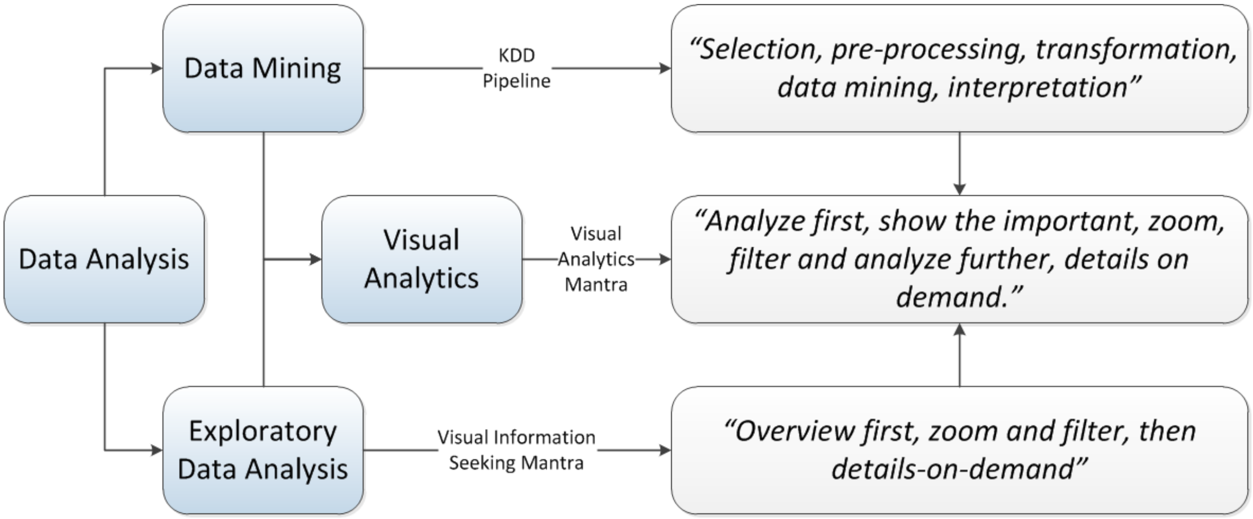

Data Mining
KDD Pipeline
"Selection, pre-processing, transformation, data mining, interpretation"
Data Analysis
Visual Analytics
Visual Analytics Mantra
"Analyze first, show the important, zoom, filter and analyze further, details on demand."
Exploratory Data Analysis
Visual Information Seeking Mantra
"Overview first, zoom and filter, then details-on-demand"

Figure 1: Data Analysis Workflows

Data mining is mainly concerned with data modeling for predictive purposes, relying heavily on automated analysis methods that can handle large amounts of data. Data mining is one of the key steps in the *Knowledge Discovery in Databases* (KDD) process.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. [Fayyad1996] proposes a systematic approach to the KDD process as a series of steps, called the KDD pipeline. The KDD pipeline is a precursor of the idea of creating a mantra to guide analysts in their data exploration journey and consists of the following steps:

- Selection
- Preprocessing
- Transformation
- Data mining
- Interpretation

A very different approach to knowledge extraction is taken by *Exploratory Data Analysis* [Tukey1977]. This discipline has a strong human component as practitioners are encouraged to visually examine their data with the assistance of summarizations and descriptive statistics. Information visualization plays a central role in exploratory data analysis; [Shneiderman1996] proposes a basic principle for the discipline in the form of the Visual Information Seeking Mantra:

"Overview first, zoom and filter, then details on demand."

Shneiderman goes on to define a task taxonomy[1] that, in addition to the four tasks in the mantra, includes three support tasks: relate, history, and extract.

The last decade has seen the disciplines of data mining and exploratory data analysis come together. Each discipline brings their strengths to bear against the challenge of an ever growing amount of data. In the new science of *Visual Analytics*, human experts use their domain knowledge and visual problem solving skills early in the data analysis process alongside the power of automated analytics. A definition for this discipline is given in [Cook2005]:

"Visual Analytics is the science of analytical reasoning supported by interactive visual interfaces."

To capture the essence of this new discipline, [Keim2009] proposed a Visual Analytics workflow that combines the KDD Pipeline and the Visual Information Seeking Mantra in an all-encompassing Visual Analytics Mantra:

"Analyze first, show the important, zoom, filter and analyze further, details on demand."

The elegant reduction of a complex workflow to a single phrase – a *mantra* – gives data explorers something to memorize and recall as needed so not to become lost in a data jungle. Now we will see how you can implement this workflow using exploration in SAS Visual Analytics.

[1] [Shneiderman1996] also introduce a taxonomy of types and in [Heer2012] the task taxonomy concept is extended to handle a broader range of interactions. These topics are not covered in this paper.

DATA EXPLORATION WITH SAS VISUAL ANALYTICS

SAS® Visual Analytics is designed from the ground up as a tool for exploratory data analysis, it also includes a number of advanced analytical tools. The the exploration capabilities, you can carry out the Visual Analytics Mantra tasks in a seamless workflow.

Figure 2 illustrates the tasks examined in this paper and how they fit together. Note that both the Visual Analytics workflow and the original Visual Information Seeking workflow are supported by our exploration capabilities in SAS Visual Anlytics. The choice of which path to follow depends largely on the characteristics of the data and on problem at hand.

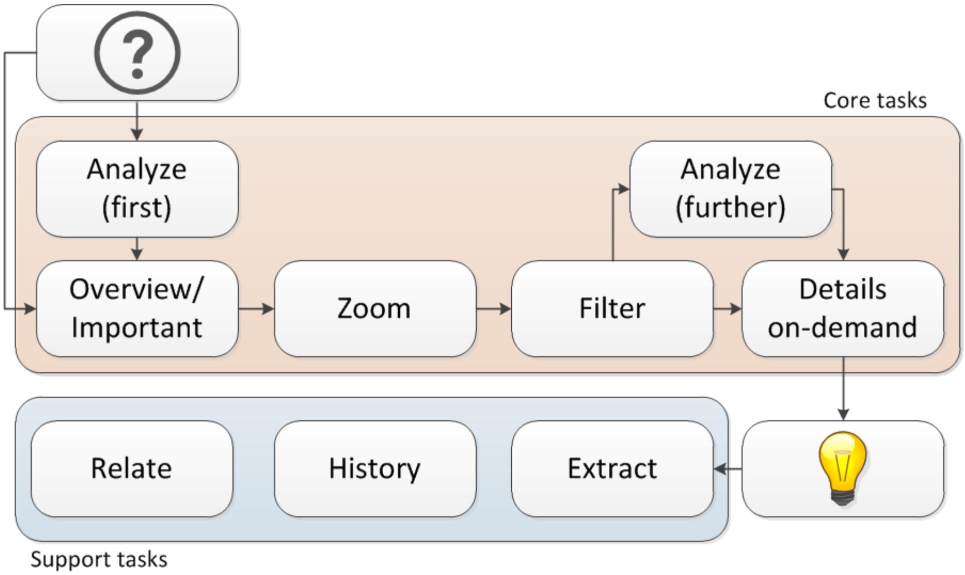


Figure 2: Visual Analytics Tasks and Workflow

CORE TASKS

We will look into the six core tasks that comprise the main workflow defined by the Visual Analytics Mantra: Analyze first, overview (show the important), zoom, filter, analyze further, and show the details.

The analytical tasks are optional, depending on the problem at hand. Complex problems that require analytics to be applied first follow the Visual Analytics Mantra. Otherwise, the data exploration reverts to the path defined in the Visual Information Seeking Mantra.

ANALYZE (FIRST)

The first task in the Visual Analytics Mantra advises you to use analytics upfront to find a starting point in your exploration. Usually, that means having a hypothesis to test, which once validated, will guide the rest of the analysis. This can be accomplished during the data preparation step, before loading your data in the in-memory server used by SAS Visual Analytics. You can also use some of the built-in analytics features to get started.

For example, you might want to find out more information about which variables influence each other. A *correlation matrix* allows you to quickly identify pairs of strongly correlated measures. If you already have an initial variable in mind and want to test which variables have (or not) a relationship to it, you can use the *correlation row*. The visualization in Figure 3 allows you to sort the variables by correlation value for a precise assessment of the relative impact of each variable. The example in Figure 3 shows a strong correlation between GDP and external debt stocks among a list of other variables with weak correlations. The strong correlation between these two variables is a good indication that this is a good place for you to focus your analysis.

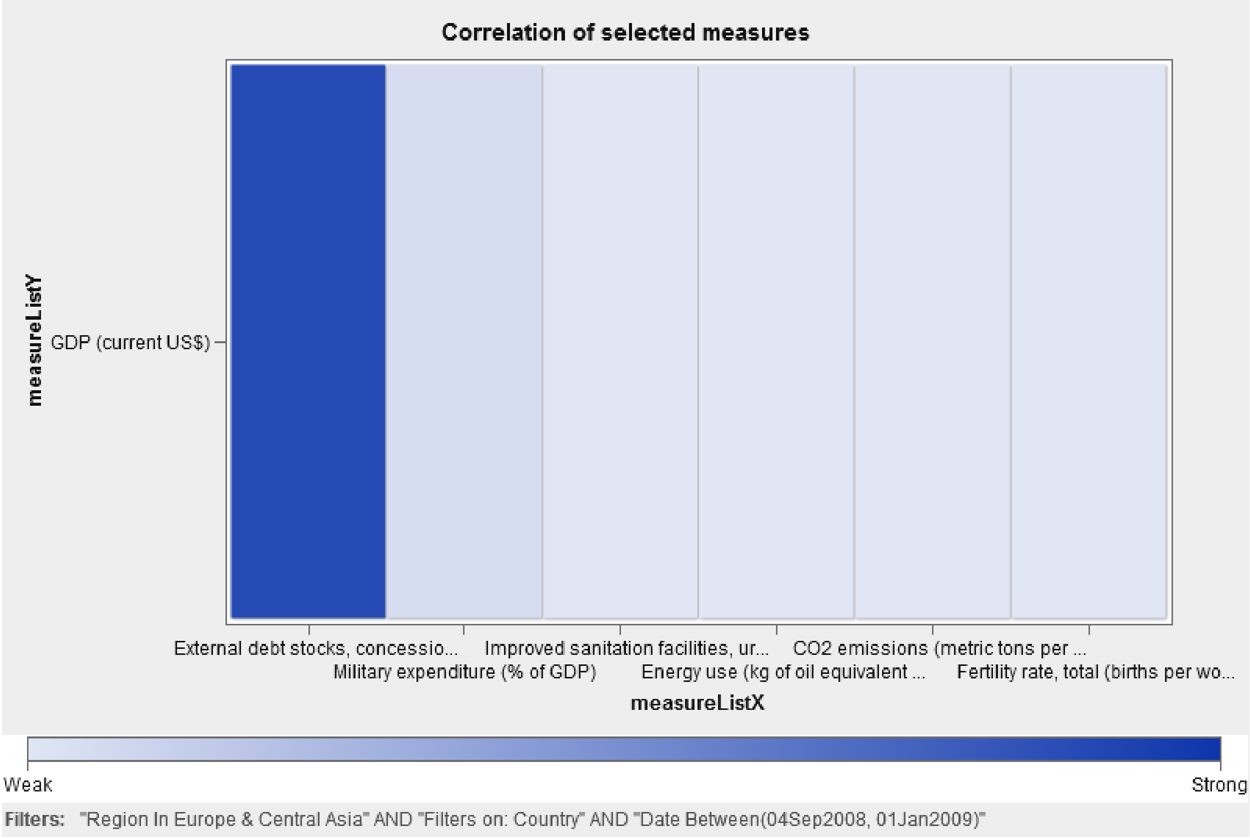


Figure 3: Correlation Row

The correlation matrix and the correlation row rely on linear regression to assess the strength of the relationship between the variables. If you suspect a relationship is not as strong as it should be, you can drill down into it to visualize the individual values in a scatter plot or in a heat map if you have a large amount of data. You can then overlay a fit line over your visualization; the fit line will tell you whether that weak linear relationship has a quadratic or cubic nature.

Another option to kick-start an exploration is a *decision tree*[2], which enables you to understand how the value of a target variable is affected (if at all) by a set of different input variables. You can create decision trees for both numerical and categorical variables. Figure 4 shows a decision tree for a numerical value. The decision tree displays the associated distributions inside each node.

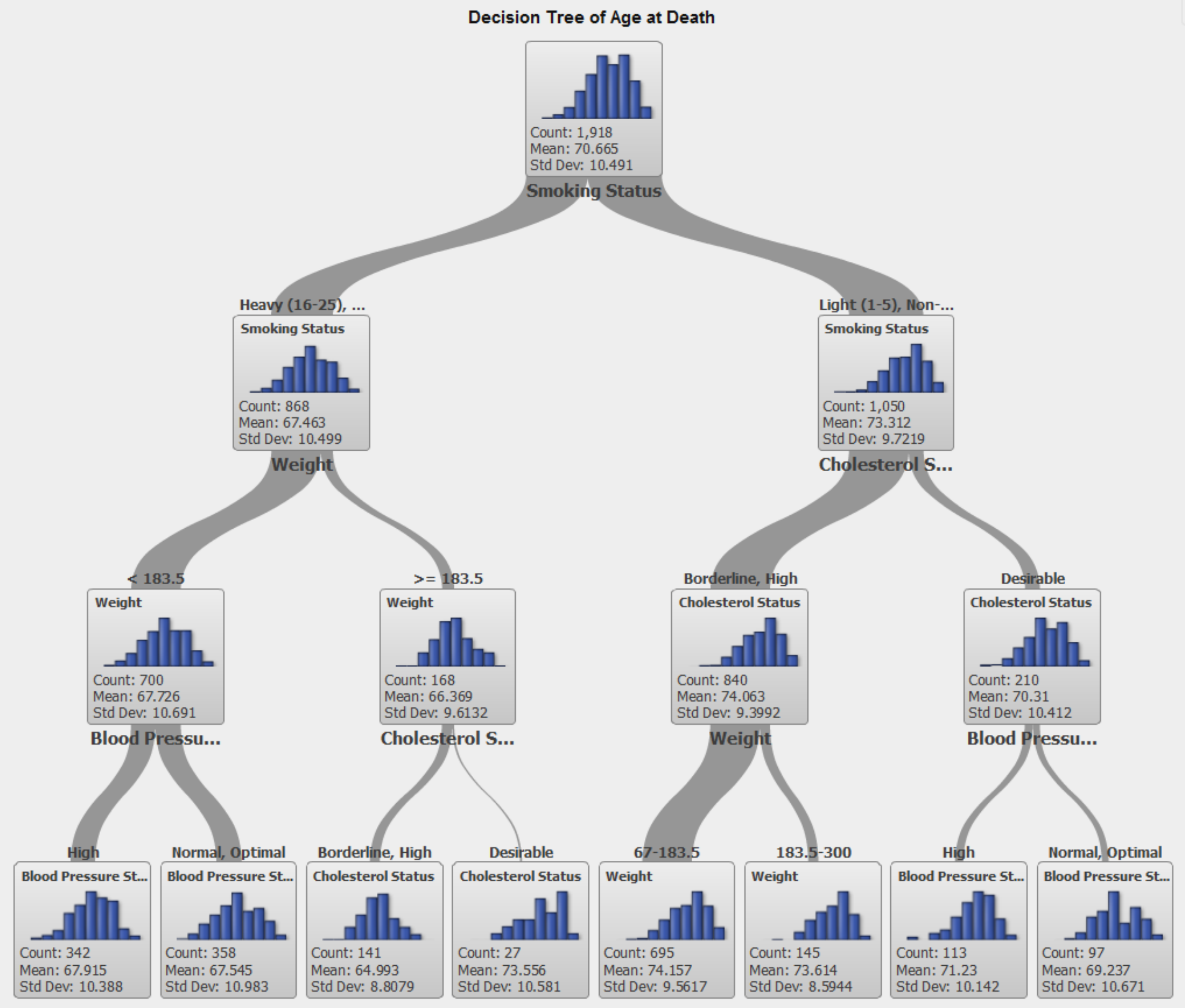


Figure 4: Decision Tree

[2] : Decision trees will be available in the 6.2 version, to be released in the second quarter of 2013.

OVERVIEW (SHOW THE IMPORTANT)

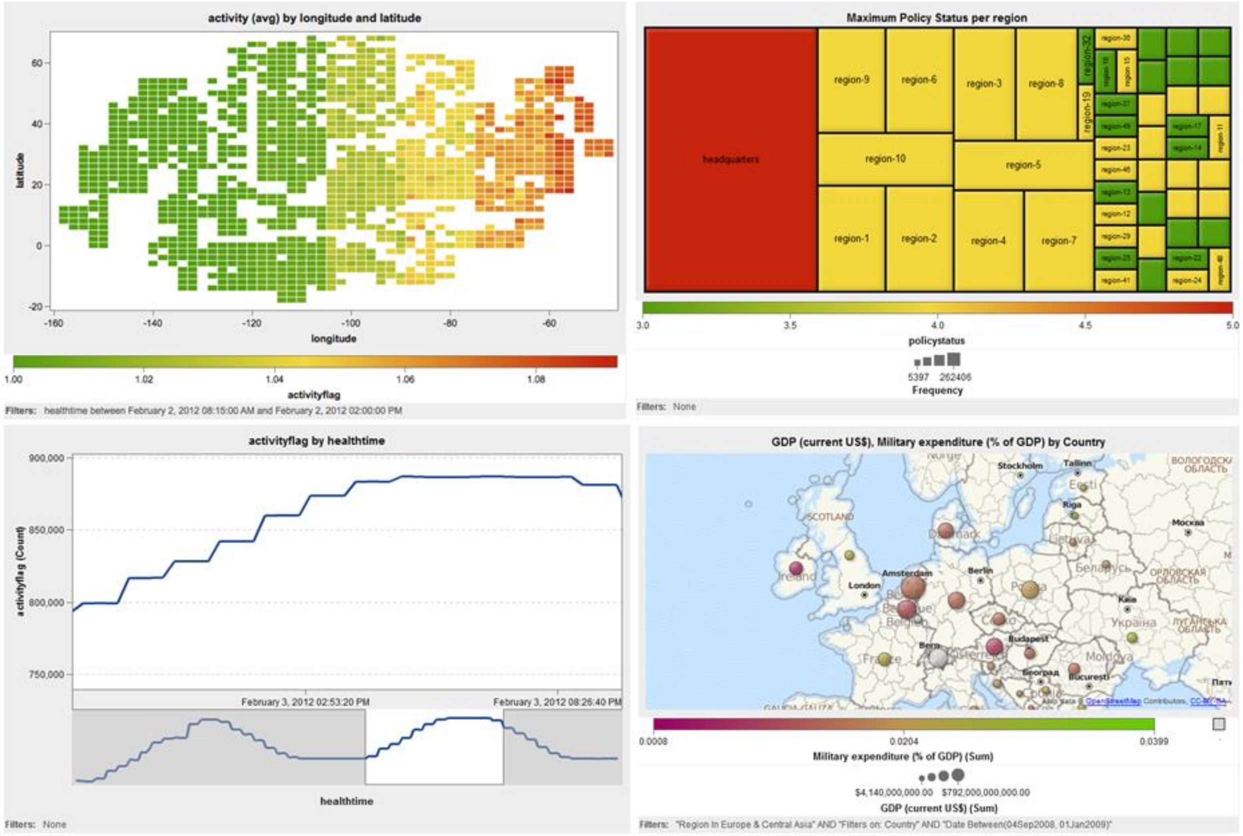
This task asks you to use your visual skills to look for patterns in a visualization that capture the essence of your data set. This task is reviewed in detail in [Few2011], who nicely explains why an overview is the starting point of a data exploration journey: "If no map exists, then you should head for high ground to get the lay of the land."

Perhaps more than in any other task, a successful overview depends on the selection of a visualization that is effective in presenting the important aspects of the data. Given the large diversity of data set types and their structure, not to mention available visualizations, selecting the right visualization for a given data set can be overwhelming.

Recognizing that challenge, the SAS Visual Analytics provides an *autocharting mode* that uses the data variables assigned to the visualization to automatically select the most suitable visualization for the data set. You can, of course, overwrite this choice at any point. During your exploration, you can change between visualization types with a single click so that you can freely experiment and focus on your exploration.

Figure 5 shows a sample of visualizations that are suitable for the initial overview of a data set. From the top left, clockwise:

- A heat map, which compares the relationships between two numeric variables and uses binning to allow the display of large amounts of data
- A treemap, suitable for the display of hierarchical data
- A geographical map so that you can visualize location-based data
- A time series with an overview pane



Filters: healthtime between February 2, 2012 08:15:00 AM and February 2, 2012 02:00:00 PM

Filters: None

Filters: None

Filters: "Region In Europe & Central Asia" AND "Filters on: Country" AND "Date Between(04Sep2008, 01Jan2009)"

Figure 5 Sample of Overview Visualizations

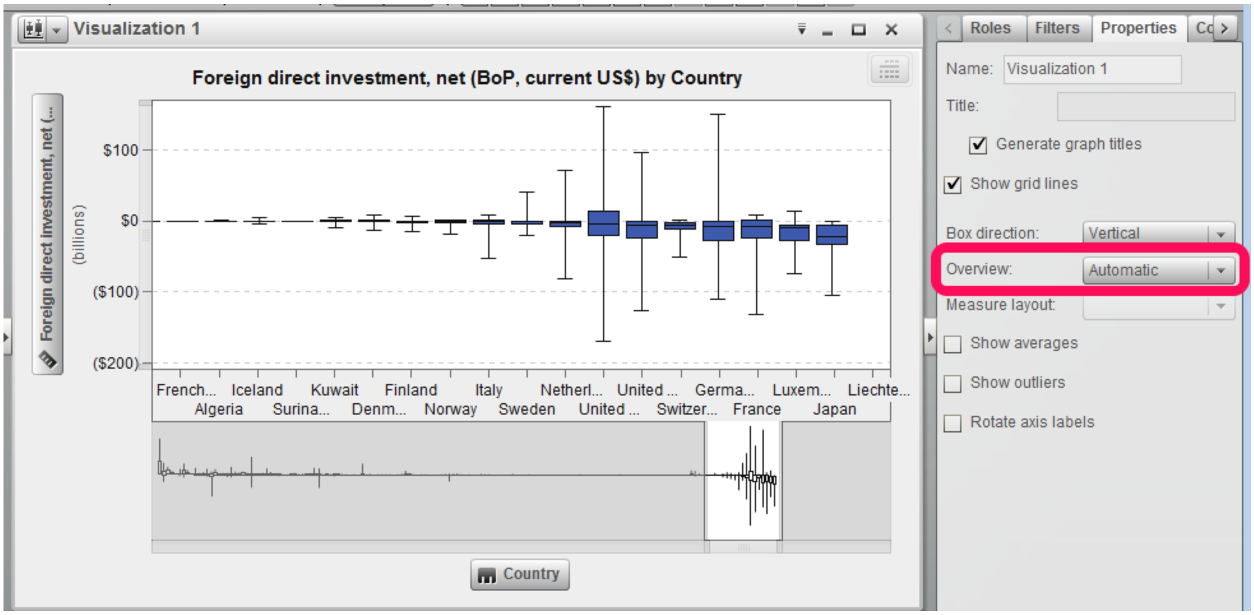
ZOOM

Once you find an area of interest, the next step is to zoom in to learn more about it. It is easy to do, using one of the multiple zoom options for visualizations.

You can hold your mouse pointer over the visualization axes to trigger the display of a zoom-scroll bar [Patel2001]. This control enables you to change the zoom level on the associated dimension as well as the area under display (panning). You can also use your mouse scroll wheel to zoom in and out of the area around the mouse pointer to display more (or less) information while preserving the current context.

Geographical maps have their own zoom control, designed specifically to provide visual feedback on the current zoom level. The zoom level controls the amount of detail displayed for the underlying location, such as roads, county and state lines.

Categorical visualizations (for example, bar and line charts, grouped box plots) support an Overview pane. The Overview pane is automatically displayed for categories with a large cardinality but can be turned on or off in the Properties tab (Figure 6). You can interact with the Overview pane to change the scope and the context of the zooming area.



Name: Visualization 1

Title:

☑ Generate graph titles

☑ Show grid lines

Box direction: Vertical

Overview: Automatic

Measure layout:

☐ Show averages

☐ Show outliers

☐ Rotate axis labels

Figure 6 Overview Pane and Properties Panel

If the data set has an underlying structure, you can take an alternative approach. By defining hierarchies and building your visualizations around them, you can drill up and down to control the amount of information that you see. A drill-down command causes the SAS Visual Analytics to replace the contents of the visualization with the breakdown of that value at the next level of its enclosing hierarchy. A breadcrumb control is added to the visualization to support navigation across hierarchies, even at previous levels (Figure 7). An animated transition ensures that you don't lose track of the context as the change takes place.

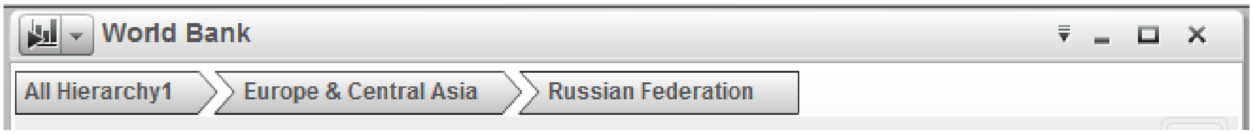


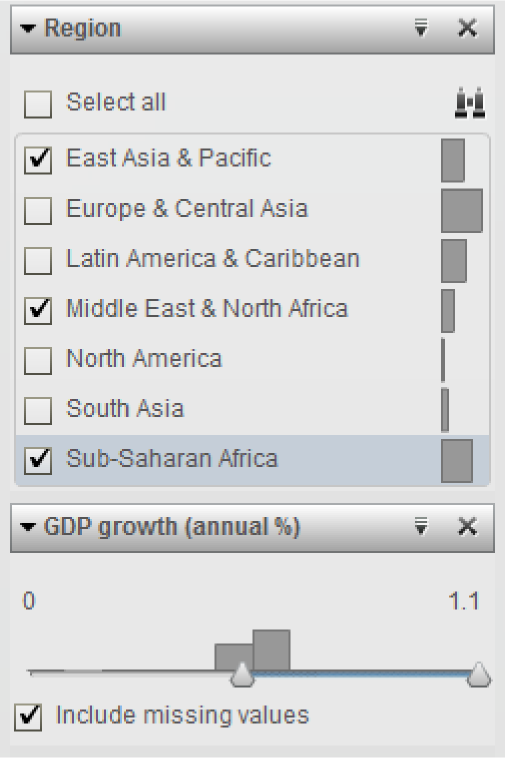
Figure 7 Breadcrumb Control.

FILTER

As negative space can be used by an artist to bring attention to an object or silence can be used by a composer to create a sense of anticipation, you too can filter out the clutter from your visualization to bring its important parts into focus. Our visual cognition channels, no matter how powerful, can be overwhelmed; knowing when to remove data from a visualization is as important as knowing when to add it.

You can drop data items directly into the Filter panel to create filters that you can then interact with to remove data in a precise way. Using the controls in the Filter panel, you can either remove an individual categorical item or groups of categorical items from the display or include only the data that falls within a specified range of time or numerical values. Local filters affect only the current visualization, whereas global filters enable you to change all visualizations with a single action.

The filter controls are implemented as *scented widgets* [Willett2007] and are decorated with bar charts that reflect the frequency of each associated item as illustrated in the example in Figure 8. Use this additional information when deciding which data to remove from your visualization.



Region

□ Select all

☑ East Asia & Pacific

□ Europe & Central Asia

□ Latin America & Caribbean

☑ Middle East & North Africa

□ North America

□ South Asia

☑ Sub-Saharan Africa

GDP growth (annual %)

☑ Include missing values

Figure 8 Filter Controls

You can also interact directly with a visualization to remove unwanted data. Just select the visual elements you want to remove or keep to the exclusion of all others, and in a couple of clicks they will be gone. A corresponding filter will be created, which you can later edit or delete. These filtering approach is particularly effective when your criteria are easier to identify visually, as when removing outliers from a box plot or from a scatter plot or when removing the top or bottom values in a bar chart.

ANALYZE (FURTHER)

At this point in the analysis, you might be close to the answer to your problem, but you might need to apply further analytics to be sure. This is a good time you test your hypothesis using analytics to decide if your work is done, if you need another iteration, or a new starting point.

For example, you analyzed, overviewed, zoomed, and filtered your way down to a segment of your market that is producing outstanding sales. Are these results uniform across the board? Or are they a reflection of a few exceptional products? Looking at the distribution of the sales values quickly surfaces the presence of outliers that you can explore further. The *box plot* visualization (displayed in Figure 9) allows you to identify outliers even for high cardinality scenarios, by visualizing them as an outlier region that you can drill into for more details.

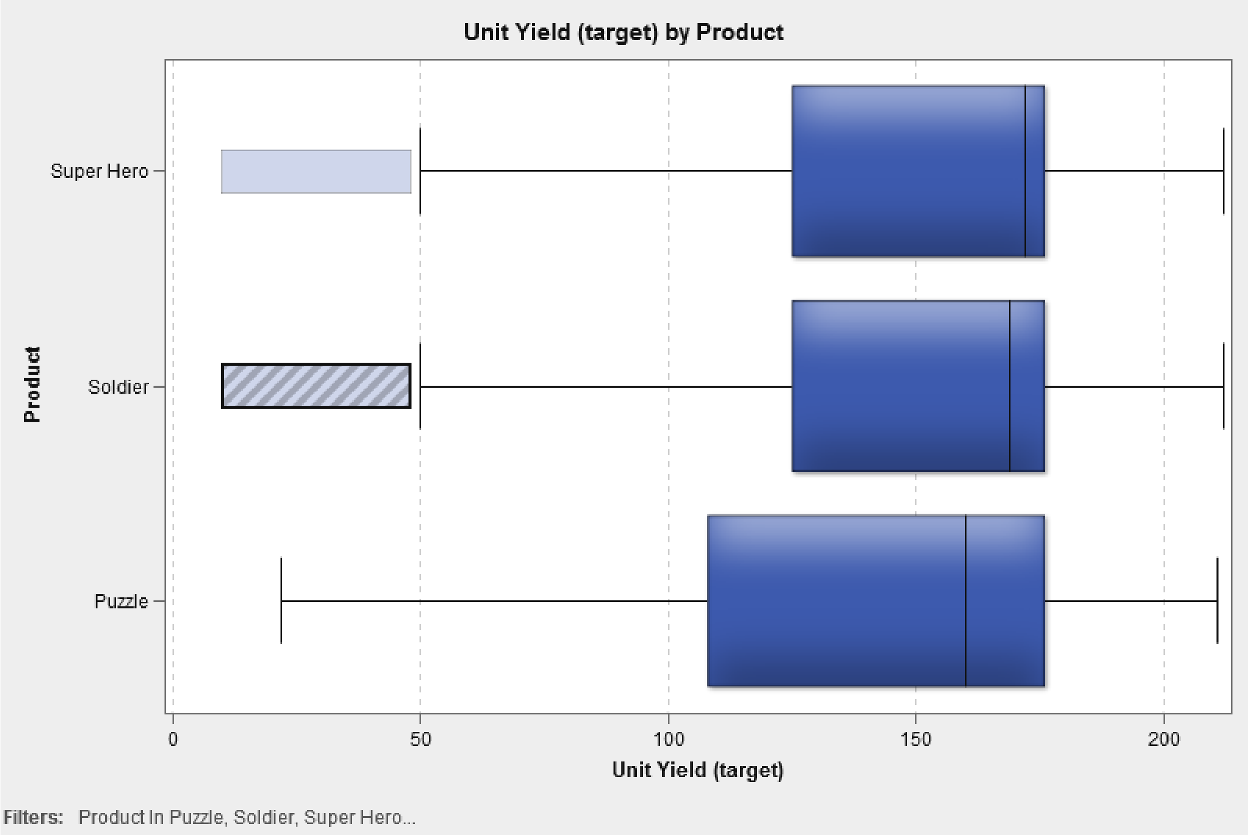


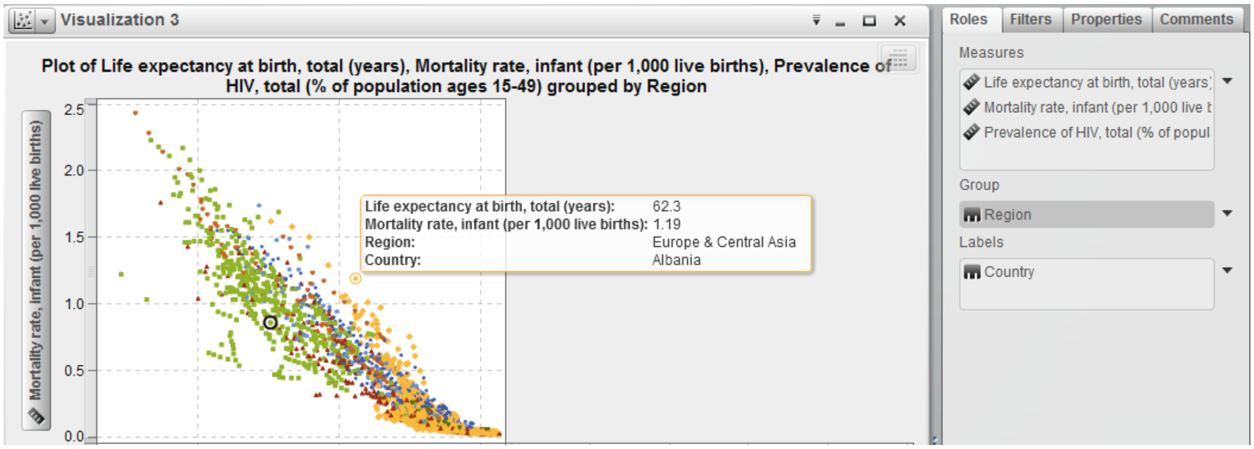
Figure 9: Box Plot with Outliers

Or maybe you are analyzing a time series and you isolated an interesting trend. You can use the SAS Visual Analytics *forecasting analysis* and *scenario analysis* to predict how it will turn out in the future. You can also find which set of variables actually influence that prediction, as well as what you would need to change to improve your odds of better results. The *scenario analysis* can provide that information.

Note that the analytics described here can also be used in the first core task, and vice-versa. It all depends on what question you're trying to answer at the corresponding step of your data analysis.

DETAILS-ON-DEMAND

Once you have narrowed your focus to a meaningful subset of the data, you should add related information to broaden your understanding and corroborate your hypothesis. For example, a scatterplot matrix displays multiple numeric values for each of your observations. With the SAS Visual Analytics, you can add additional variables to the Labels role (Figure 10) so that their values are also displayed in a data tip when you hold your mouse pointer over points of interest.



Plot of Life expectancy at birth, total (years), Mortality rate, infant (per 1,000 live births), Prevalence of HIV, total (% of population ages 15-49) grouped by Region

Life expectancy at birth, total (years):	62.3
Mortality rate, infant (per 1,000 live births):	1.19
Region:	Europe & Central Asia
Country:	Albania

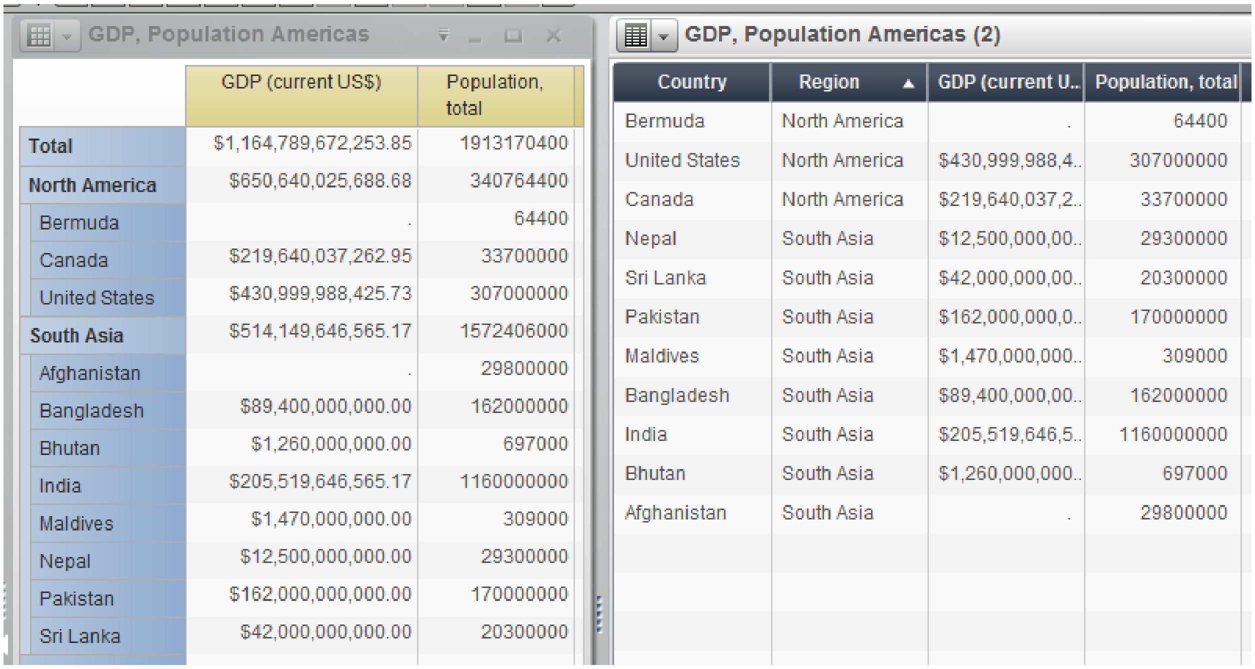
Measures
- Life expectancy at birth, total (years)
- Mortality rate, infant (per 1,000 live b
- Prevalence of HIV, total (% of popul

Group
- Region

Labels
- Country

Figure 10: Scatter Plot with Label

All SAS Visual Analytics visualizations have an associated table view that displays the (possibly aggregated) data behind the visualization in detail. You can also display the raw data in a table visualization, or see it aggregated in a crosstab. Both visualizations (Figure 11) are useful when you want to add more variables to expand the current context.



GDP, Population Americas

	GDP (current US$)	Population, total
Total	$1,164,789,672,253.85	1913170400
North America	$650,640,025,688.68	340764400
Bermuda	.	64400
Canada	$219,640,037,262.95	33700000
United States	$430,999,988,425.73	307000000
South Asia	$514,149,646,565.17	1572406000
Afghanistan	.	29800000
Bangladesh	$89,400,000,000.00	162000000
Bhutan	$1,260,000,000.00	697000
India	$205,519,646,565.17	1160000000
Maldives	$1,470,000,000.00	309000
Nepal	$12,500,000,000.00	29300000
Pakistan	$162,000,000,000.00	170000000
Sri Lanka	$42,000,000,000.00	20300000

GDP, Population Americas (2)

Country	Region ▲	GDP (current U..	Population, total
Bermuda	North America	.	64400
United States	North America	$430,999,988,4..	307000000
Canada	North America	$219,640,037,2..	33700000
Nepal	South Asia	$12,500,000,00..	29300000
Sri Lanka	South Asia	$42,000,000,00..	20300000
Pakistan	South Asia	$162,000,000,0..	170000000
Maldives	South Asia	$1,470,000,000..	309000
Bangladesh	South Asia	$89,400,000,00..	162000000
India	South Asia	$205,519,646,5..	1160000000
Bhutan	South Asia	$1,260,000,000..	697000
Afghanistan	South Asia	.	29800000

Figure 11: Crosstab and Table

The dimensionality of your data also plays a role here. The hierarchies defined in the Zoom task also provide a great way of gathering details-on-demand. Hierarchies allow you to drill down to see more details in bubble plots, box plots, bar charts, heat maps and geo maps.

SUPPORT TASKS

The support tasks defined in the Visual Information Seeking Mantra enhance the basic functionality offered by analytics and visualizations, and enable you to reach the end of your exploration more quickly. We will take a closer look and then and how they are supported by with SAS Visual Analytics.

RELATE

While exploring the data, you often look at its different facets at the same time. It is very common that the answer to your exploratory mission is revealed not by a single visualization but by the relationships among multiple visualizations. These relationships can take many forms, such as comparisons, linked selections, and calculated columns.

Comparisons

Critical aspects of the data are revealed in the similarities and differences exposed by comparing variations of the same visualization.

Comparisons for position and size are facilitated by the flexible arrangement of visualizations in the display and by a variety of layouts for visualizations based on multiple categories and measures. Among them is the support for a common axis that is shared across related visualizations, using the *small multiples* [Tufte1990] approach (illustrated in Figure 12). This is particularly useful when comparing measures of different magnitude or format, with the drawbacks associated with dual-axis charts.

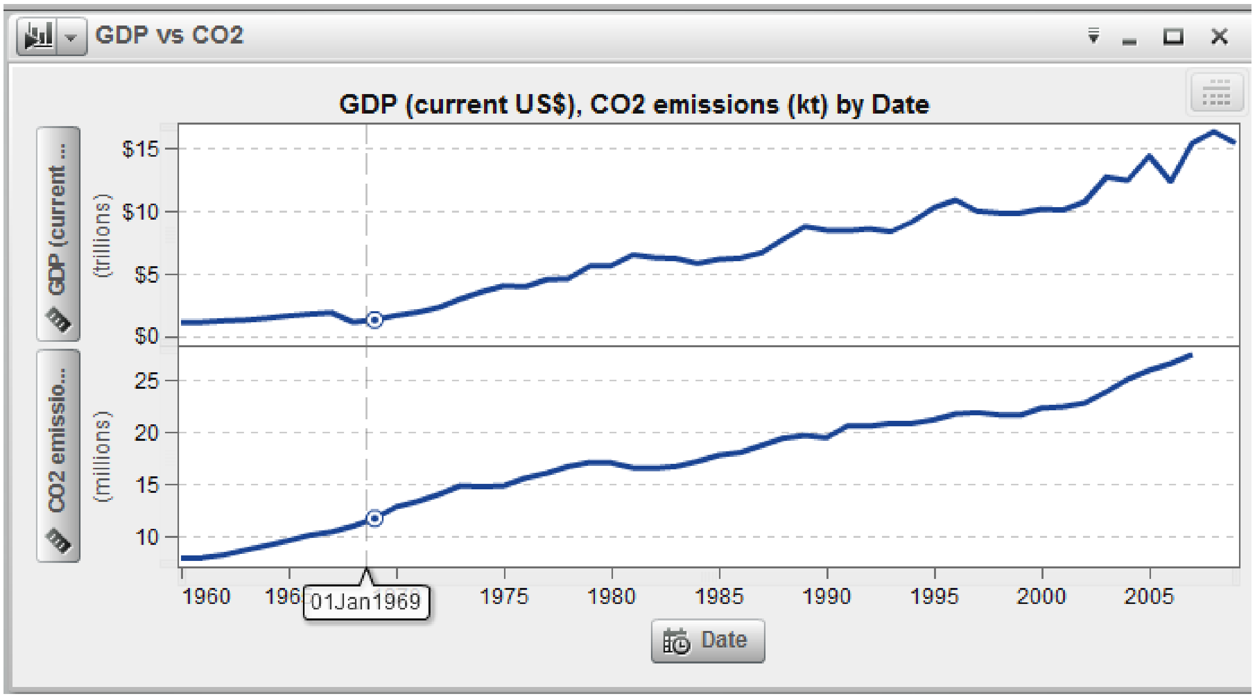


Figure 12: Small Multiples

Color is a common way to express values in visualizations. For example you might be looking at a geo map where color is used to express the number of accidents per city, and you want to compare it with a similar geo map for a different state. In the SAS Visual Analytics, effective comparison of values mapped to colors is enabled through:

- The ability to enforce the same color mapping across visualizations
- The ability to copy color mappings between visualizations.

Linked Selection

Each view of your data can reveal interesting features. *Linked selection* makes it possible to build a more complete mental model of a highlighted feature by seeing how it is displayed in other views. This makes apparent the relationships between the selected feature and other features in the data set [Ward2010].

SAS Visual Analytics supports not only linked selection but also *linked brushing*, where you can interactively change your selection and see the corresponding selection changing in real time in the other views. Linked brushing is one of the most powerful interactive tools for doing exploratory data analysis through visualizations [Few2011].

Calculated Columns

Like pieces in a puzzle, some patterns only reveal themselves when their parts come together. With SAS Visual Analytics, you can create *calculated columns* to define expressions based on multiple variables and pre-defined functions.

For example, you might have a table with financial results and public information about different companies. You have columns for price and earnings and would like to find out which company would be a better prospect for investment. Given the differences related to company size and industry, just looking at the raw numbers does not allow for an effective comparison; you need to first transform the data to extract the essence hidden within.

Simply create a new calculated column for the P/E ratio (price/earnings) and you will have a much better perspective of the data. You can then check how it changes over time or how it relates to other variables.

HISTORY

Sometimes the path is as important as the destination. An exploration will almost never proceed in a straight line from problem to insight, and there is much to be learned in the detours along the way. So it is important that you have the freedom to explore without regrets.

With full support for undo and redo operations, you can change your mind as many times as you want.

Sometimes the problem is that you can't make up your mind. Should you look at your data as a bar chart to focus on the values or as a geo map to concentrate on the location? Should you look for patterns on the North region first, or the South? When you come to these forks, you don't have to take the road less traveled [Frost1920]. Use the duplicate option to have both versions of your visualization available at once.Duplicate, make the necessary changes and compare the results; you are likely to find something new.

EXTRACT

No analyst is an island; eventually, you will need to discuss your exploration with a co-worker. SAS Visual Analytics lets you send e-mails to your co-workers with a link to your exploration. This enables your team to collaborate on the analysis, recording thoughts and impressions in context by using comments.

When it is time to share your conclusions with your boss or with your customers, you can use the following strategies:

- Save individual visualizations as images for inclusion in presentations or documents.
- Share a subset of your visualizations as a PDF file.
- Export your exploration as a SAS Visual Analytics report.

SAS Visual Analytics reports are an ideal vehicle to share your results as they can be further refined visually and displayed on mobile devices or on the web.

If you just want to capture your results as data, visualization summary tables can be exported as CSV files.

CONCLUSION

We examined how data analysis has evolved to deal with the challenges posed by big data and how some of its disciplines have merged in the field of visual analytics, with a corresponding merge of basic workflows.

We looked at how these basic workflows are composed of a sequence of steps that are captured in easy to remember mantras.

Finally we showed how the SAS Visual AnalyticsExplorer enables you to execute core and support tasks in the Visual Analytics workflow.

So don’t forget your mantras and be confident that the next time you descend into the data jungle, you will be able to see it all – both the forests and the trees.

REFERENCES

- [Abousalh2012] Abousalh-Neto, Nascif; Kazgan, Sumeyye. *Big data exploration through visual analytics* 2012 IEEE Conference on Visual Analytics Science and Technology (VAST) Issue Date: October 2012
- [Cook2005] Thomas, J., Cook, K. 2005. *Illuminating the Path: Research and Development Agenda for Visual Analytics*. IEEE-Press.
- [Dumbill2012] Dumbill, Edd. 2012.*Volume, Velocity, Variety: What You Need to Know About Big Data* http://www.forbes.com/sites/oreillymedia/2012/01/19/volume-velocity-variety-what-you-need-to-know-about-big-data/
- [Heer2012] Heer, Heffrey; Shneiderman, Ben. 2102.] Interactive Dynamics for Visual Analysis 2012 http://queue.acm.org/detail.cfm?id=2146416
- [Keim2009] Keim, Daniel A.; Mansmann, Florian; Thomas, Jim. 2009. *Visual analytics: How much visualization and how much analytics*. SigKDD Explorations Journal
- [Fayyad1996] Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy. 1996. “From Data Mining to Knowledge Discovery: An Overview". *Advances in Knowledge Discovery and Data Mining,* pp.1-34. Menlo Park, CA, AAAI Press / The MIT Press
- [Few2011] Few, Stephen. 2011. Exploratory Vistas - Ways to Become Acquainted with a Data Set for the First Time, Perceptual Edge Visual Business Intelligence Newsletter July/August/September 2011
- [Frost1920] Frost, Robert. *The Road Not Taken*
- [Patel2001] Patel, Himesh; Matange, Sanjay. *New Visualization in V8.2*. Proceedings SUGI 2001
- [Shneiderman1996] Shneiderman, Ben. 1996. *The eyes have it: A task by data type taxonomy for information visualizations.* In IEEE Symposium on Visual Languages, pages 336-343.
- [Tukey1977] Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading MA: Addison-Wesley
- [Tufte1990] Tufte, Edward. 1990. *Envisioning information.* Graphics Press Cheshire, CT, USA
- [Ward2010] Ward, Matthew; Grinstein, Georges; Keim, Daniel. 2010. *Interactive Data Visualization – Foundations, Techniques, and Applications*
- [Willett2007] Willett, Wesley; Heer, Jeffrey; Agrawala, Maneesh. November 2007. *Scented Widgets: Improving Navigation Cues with Embedded Visualizations.* IEEE Transactions on Visualization and Computer Graphics archive, volume 13 Issue 6, pages 1129-1136

RECOMMENDED READING

- *SAS® Visual Analytics 6.1 User’s Guide (http://support.sas.com/documentation/cdl/en/vaug/65747/PDF/default/vaug.pdf)*
- *Now You See It: Simple Visualization Techniques for Quantitative Analysis, Stephen Few, 2009*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nascif Abousalh-Neto
SAS Institute Inc
SAS Campus Drive S4082
Cary, NC 27513-2414, USA
(919) 531-0123
nascif.abousalh-neto@sas.com
http://www.linkedin.com/in/nascif